基于 ChatGPT 的情绪稳定性计算机自适应题库开发的探索

高垚杰[2] 齐运晓[2] 马苑秋[2] 刘拓[123]（通讯作者）

（1 教育部人文社会科学重点研究基地天津师范大学心理与行为研究院，天津，300387）

（2 天津师范大学心理学部，天津，300387）

（3 学生心理发展与学习天津市高校社会科学实验室，天津，300387）

摘　要 为得到一个质量良好的大型题库，进行传统形式项目开发所耗费的大量人力物力制约着目前计算机化自适应测验的发展与运用，而基于最新自然语言处理技术的自动项目生成有望解决这一难题。随着基于 Transformer 架构的生成式预训练模型的进步，根据特定测量目标（尤其是非认知任务），自动生成测验项目并以此为基础建立计算机自适应题库成为可能。本研究旨在利用最新版本的 ChatGPT 生成大量中文版测量情绪稳定性的人格项目，通过单维性检验、IRT 模型选择、项目分析、题库质量分析等题库构建步骤以及模拟的计算机化自适应测验，探索这些项目对于计算机化自适应测验的适用性并与已被广泛使用的情绪稳定性项目进行性能对比，最终形成了一个质量良好的情绪稳定性题库。

关键词 计算机化自适应测验，自动项目生成，自然语言处理，情绪稳定性，项目反应理论

通讯作者: 刘拓，E-mail: mikebonita@163.com

Exploration of Computerized Adaptive Item Bank Development for Emotional Stability Based on ChatGPT

Gao Yaojie[2] Qi Yunxiao[2] Ma Yuanqiu[2] Liu Tuo[1,2,3] (Corresponding author)

（1 Key Research Base of Humanities and Social Sciences of the Ministry of Education, Academy of Psychology and Behavior, Tianjin Normal University, Tianjin 300387）

（2 Faculty of Psychology, Tianjin Normal University, Tianjin 300387）

（3 Tianjin Social Science Laboratory of Students' Mental Development and Learning, Tianjin 300387）

Abstract

To obtain a high-quality large-scale item bank, the extensive manpower and resources required for traditional project development have been constraining the development and application of computerized adaptive testing. However, the automatic item generation, based on the latest natural language processing technology holds promise in addressing this challenge. With the advancements in generative pre-trained models based on the Transformer architecture, the generation of items tailored to specific measurement objectives (especially non-cognitive tasks) becomes feasible. This study aimed to utilize ChatGPT to generate a large number of Chinese version personality items measuring emotional stability and to establish a computerized adaptive item bank based on this premise.

We utilized ChatGPT based on GPT-4 Turbo to generate 114 items measuring emotional stability. Following expert review, 75 items were retained and formed the GPT item bank, while 42 widely-used items were selected to form the classic item bank. Testing was conducted on the aforementioned items, yielding 479 valid participants. Additionally, sample data from two separately administered measures, CBF-PI-B and BFI-2, were going to be used for subsequent cross-sample reliability comparisons. Procedures for item bank construction including unidimensionality test, IRT model selection, item analysis, and item bank quality analysis, as well as simulated computerized adaptive testing, were employed to assess the quality and CAT performance of the item bank.

After the above analysis steps, it was found that all items in the classic item bank and the GPT item bank passed the unidimensionality test, showing no differential item functioning, and had good

Corresponding author: Liu Tuo, E-mail: mikebonita@163.com

discrimination parameters and reasonable difficulty distribution. Both item banks provided high test information and marginal reliability for most trait levels of the examinees, with low measurement error. The overall item bank formed by combining all items remained of good quality. Simulation results of computerized adaptive testing showed that all three item banks achieved high validity with fewer items compared to traditional tests for the same level of precision. Under the same testing length, GPT item bank exhibited higher reliability and demonstrated stability across samples. Additionally, comparison revealed that the CAT performance of the GPT item bank even exceeded that of the classic item bank, while the overall item bank performance was slightly better than that of the GPT item bank.

This study innovatively explores the development of a computerized adaptive item bank using the latest version of ChatGPT, validating the feasibility of this user-friendly project generation tool. Through comparison with previous research results, it reconfirms the excellent quality of projects generated by GPT-4. The study showcases the immense potential and possibilities of large language models in project development, particularly in the creation of large-scale item banks, while also indicating at a shift in the responsibilities of psychologists in future project development.

Keywords Computerized Adaptive Test, Automatic Item Generation, Natural Language Processing, emotional stability, Item Response Theory

1 前言

计算机化自适应测验(Computerized Adaptive Test, CAT)是一种基于计算机的测试方法，可以在测试过程中根据受测者的作答行为匹配出最适合其作答的项目，从而提高测试的效率和得分的有效性(Fliege et al., 2005; Jiao & Lissitz, 2020)。CAT 相较于基于经典测量理论(Classical Test Theory, CTT)的传统测验，一方面使受测者不再需要作答完量表的所有项目，在提高测量效率的同时减少了其认知负担，避免了因长时间测试造成的无聊感而导致测验准确性降低；另一方面实现了“因材施测”，减少了因项目难度与受测者特质水平相差太大而造成的测量误差。即使每位受测者测验项目不同，依然能够实现作答结果的可比性，则是因为 CAT 以项目反应理论(Item Response Theory, IRT)为基础。在 IRT 中，只要所有项目的参数标定在同一个量尺上，那么即使作答的项目数量和难度不同，根据作答结果估计出的每位受测者的能力值（特质水平）也是可比较的。因此，由存在共同测量目标且参数处于同一尺度上的项目组成的题库便是 CAT 的基本前提，而为了实现这点，既要求开发者在初期收集到大量项目并进行大规模施测来获得稳定的项目参数，以此构建最初的题库，又需要在后续使用过程中，定期管理和补充题库，如控制曝光率过高的项目，淘汰质量欠佳的项目等。由于传统的项目开发往往依赖于相关领域的专家，且需要耗费大量的人力物力与财力，因此建立一个包含丰富项目且能不断更新的题库是制约目前 CAT 发展与运用的一大重要瓶颈(Gierl & Haladyna, 2013; Gierl & Lai, 2018)。

为了解决这一难题，自动项目生成(Automatic Item Generation, AIG)与 CAT 的结合被认为有非常广阔的前景(Hommel et al., 2022)。Gierl 与 Lai(2018)将 AIG 总结为 3 个步骤，首先由测验主题领域的专家(subject-matter expert, SME)使用认知模型（一种强调解决特定领域内问题所需的知识、技能和能力的表征）组织和构建项目生成所需的内容；第二步是由专家开发项目的模板，该模板明确了认知模型中的内容可以放置的位置，使得能够通过替换模板中该位置的具体内容来生成项目；最后则是通过计算机算法，将认知模型的内容放置到模板当中，以及通过替换内容来生成新项目。目前 AIG 已被大量研究者运用到教育背景下的测验开发中(Gierl et al., 2012; Gierl & Haladyna, 2013; Lai et al., 2016; Kurdi et al., 2020)，尽管这种基于模板的 AIG 可以生成大量认知测试的项目，但并不太适用于非认知的测验，如心理测验中常见的人格测试，这些测试的项目往往有更复杂的语义、情境和细微的差别(Hernandez & Nie, 2022; Hommel et al., 2022; Lee et al., 2023)。以 BFI-2 中文版(Zhang et al.,

2022)中神经质维度的“我是一个能够控制自己的情绪的人”这一项目为例，如果“控制”或“情绪”被视为模板中可替换的词，那么有意义的替换对象屈指可数，且采用同义词或近义词替换后产生的新项目的作用十分有限，并限制了整个量表从不同情境下测量神经质的能力。所幸的是，随着自然语言处理(natural language processing, NLP)技术的发展，不断有研究者尝试运用算法来自动生成非认知的项目(Von Davier, 2018; Hernandez & Nie, 2022; Hommel et al., 2022; Götz et al., 2023; Lee et al., 2023)。

NLP 是人工智能(Artificial Intelligence, AI)里通过开发量化模型来让计算机理解、分析和生成人类语言的一个子领域(Goldberg, 2017; Lee et al., 2023)。随着神经网络（一种模拟生物神经网络的计算模型，能将输入的数据在网络中根据给定的映射进行连续的转化后再输出；Goldberg, 2017）的引入，NLP 的语言模型取得了巨大的进展(Götz et al., 2023)。Von(2018)便开创性地将基于当时最先进的长短期记忆(long short-term memory, LSTM)网络的语言模型应用于非认知项目的 AIG，但该技术只能模仿示例项目的语法结构，无法根据特定的目标概念生成项目，并存在计算量极大、难以始终保持项目语法的正确性等其他局限(Hernandez & Nie, 2022; Hommel et al., 2022)。此外，本质上 LSTM 是一种基于监督机器学习的模型，即通过有标记的输入（例如一部分邮件被标记为垃圾邮件，而其他不是）进行训练，以此使模型逐渐学习如何正确预测新的输入的标签，因此需要大量有人工标注的样本数据进行模型训练，然而大量良好的样本项目在实际中往往是难以获得的，这也正是使用 AIG 的原因之一(Goldberg, 2017; Lee et al., 2023)。因 OpenAI 发布的基于 Transformer 架构的生成式预训练模型(Generative Pre-trained, GPT)及其迭代版本展现出了优越性能，使其受到了利用 AIG 生成非认知项目的研究者的青睐。

自注意力机制是 Transformer 一大特点，该机制能将一个输入序列（例如一句话）中的任意单词与序列中其他单词进行交互，从而使每个单词的输出包含了与其他单词间关系的信息， 这让 Transformer 不再需要同先前的循环模型（如 LSTM）一样储存该单词前的所有单词，而是能通过数学计算把握单词间的微妙联系，通过关键的上下文信息来预测内容（具体的数学原理可以参见 Vaswani et al., 2017）；另一方面 Transformer 可以并行处理文本序列中的不同单词，相较于循环模型，减少了计算量，让训练更大、性能更强的模型成为可能(Götz et al., 2023)。GPT 正是以 Transformer 的解码器为基础开发而来的，即一种自回归模型，仅使用序列中前 n 个单词的信息（此时第 n 个单词后的信息被遮掩了）预测第 n+1 个单词，接着将文本中实际的 n+1 个单词与预测的单词进行比对从而训练模型，简单来说

就是通过文本中的上文信息来预测下文信息以此循环直至完成整个文本序列。这种自回归形式已被证明相较于同时通过上下文信息来推断中间某个单词的“完形填空”形式，更适合应用于文本生成领域(Hommel et al., 2022)。作为预训练模型，GPT-3 包含 1750 亿个参数，并在海量文本数据中进行了训练(Brown et al., 2020)，而使用者可以仅通过几个例子（Few-Shot，一般取 10-100）来让模型执行特定的任务（例如生成某种人格的测验项目），不再需要大量人工标注的训练数据或针对特定任务对模型进行微调。后续 OpenAI 发布的基于 GPT-3.5 及 GPT-4 的 ChatGPT，因其仅依靠对话而无需任何代码的使用形式，受到了广泛的关注。尽管更多技术细节没有公布，但通过 GPT-4 在各项任务中取得的优异成绩及对比先前版本取得的显著进步，尤其是其在 AP 心理学测验中取得了 5 分的成绩(Achiam et al.,2023)，使其生成非认知项目的表现令人期待。此外，GPT-4 同样有优秀的理解与生成中文的能力，鉴于国内相关研究的空白，验证 GPT 能否在中文语境下生成质量良好的项目同样必要。

因此本研究旨在基于最新版本 ChatGPT 生成中文版的人格项目，并初步探索使用这些项目构建题库进行 CAT 的可能性并检验其性能，以期最终建立一个质量良好的计算机自适应题库。鉴于相关领域探索仍处于起步阶段，本研究将参考前人范式，继续以人格特质中最被广泛使用的大五人格(John et al.,2008)作为项目生成目标，又由于经典的 CAT 需满足单一的测量目标，因此将生成情绪稳定性（emotional stability，简称 ES）维度的项目确认为最终的生成任务。一方面因为 ES 受到了社会越来越广泛的关注，另一方面其在组织管理、心理健康、学校教育、决策行为等方面都有重要作用，被认为是心理健康的最重要的预测指标(Bajaj et al., 2019; Bec & Becken, 2021; Margetić et al., 2022; Park et al., 2022; Wettstein et al., 2021)。

2 方法

2.1 测量工具

为了让 ChatGPT 生成符合要求的项目，我们使用中文进行了一系列的提示工程(prompt engineering)：（1）让 GPT 扮演经验丰富的心理学家；（2）为其介绍国内外知名学者对于情绪稳定性的定义；（3）描述 ES 与大五人格模型的关系，并以此介绍 ES 作为一种人格特质常常通过大五人格测验中的神经质子维度进行测量，由此向 GPT 提出要参考已有项目来制定全新的测量项目这一任务；（4）为 GPT 详细规定制定项目过程中必须遵循的基本原则，包括原创性（避免语义过于相似）、避免重复（新项目间及与参考项目间的重复）、明确测

量目标（情绪稳定性）、丰富的项目形式（避免项目在结构上过于相近）、多样性与全面性等。尤其最后一点值得注意，该原则源自前人研究提出的 GPT 生成项目存在情境上过于相似的问题(Lee et al., 2023)，因此在阐述该规则时，研究者应当为 GPT 提供与测验目标相关的尽可能多的情境及相关因素作为辅助，这些解释有助于提高生成项目的质量(Lampinen et al., 2022)；（5）最后，为 GPT 逐次提供不同的大五人格问卷中测量 ES 的项目、指导语与计分方式作为参考（而非模仿模板），并以此要求其制定出等同数量的项目（含计分方式）。在这个过程中，如若 GPT 生成了明显不符合制定规则的项目，如完全照搬了参考项目或与先前制定的项目重复，则可以提示 GPT 再次回顾制定项目必须遵守的基本原则，并完善或重新制定该批次项目。

根据上述步骤，我们得到了 GPT 制定的 114 道测量情绪稳定性的项目，然后邀请 10 位心理学专业的研究生对项目的语法与内容有效性进行判断。根据结果，首先剔除了 1 道表述不当的项目，接着关于内容有效性，10 位专家对每个项目代表情绪稳定性这一人格特质的程度进行 4 点评分。根据评分结果，计算了修正后的 kappa 系数(k^*; Polit et al., 2007)，按照 Polit 等人的评估标准，一般认为 k^*大于 0.74 的项目属于优质项目，因此所有不符合该标准的项目被剔除，最终保留了 75 道项目以进行正式施测。值得一提的是，这 75 道项目没有经过任何人为的修改或编辑，纯粹是由 GPT 生成的。

另外我们选择了 4 个已被广泛使用的大五人格量表中共计 42 道测量情绪稳定性（神经质）的项目加入到施测项目中。其中包含中国大五人格问卷简式版的 8 题（CBF-PI-B; 王孟成等，2011），BFI-2 中文版的 12 题(Zhang et al., 2022)，翻译后的 TIPI-10 的 2 题(Gosling et al., 2003)与 IPIP-BFAS 的 20 题(DeYoung et al., 2007)。在本研究中，这四个量表的情绪稳定性维度的 Cronbach's α 分别为 0.921，0.935，0.799， 0.955。

题库中的项目共包含多种计分方式，每个项目保持原有计分方式不变。

2.2 被试

本研究包含 3 个样本数据。主要的样本 1 是对 117 道情绪稳定性项目进行了方便抽样，共获得有效数据 479 人（其中男性 163 人），平均年龄为 22.82 岁(SD=6.52)。另外，本研究还分别单独施测了 CBF-PI-B 与 BFI-2 中文版的情绪稳定性项目以进行跨样本的信度比较，获得的有效样本分别记为样本 2 与样本 3。样本 2 共包含 2484 人（男性 820 人），样本 3 共包含平均年龄为 28.58（SD=10.58）的 655 人（男性 197 人）。

2.3 分析方法

在进行 CAT 前，首先需要得到质量良好的题库。为方便表述，本研究将已有量表的 42 个项目组成的题库记为经典题库，GPT 生成的 75 个项目记为 GPT 题库，并分别对两个题库在样本 1 上的数据进行主成分分析与单维性检验、IRT 模型选择、项目分析以及题库整体信息量与边际信度的分析，以此对两个题库的质量进行检验与比较，筛除掉不符合要求的项目。再将最终保留的所有项目结合形成总题库，并按照上述步骤再次对其质量进行检验，以期获得一个更丰富的终版题库，并验证两种来源的项目组成题库的可结合性。之后，本研究将最终保留的 3 个题库进行模拟 CAT，一方面进一步比较 GPT 制定的项目与经典项目的性能优劣，另一方面验证 CAT 相较于传统测验的性能提升。具体步骤如下：

2.3.1 主成分分析与单维性检验

首先为保证题库的质量，与测量目标相关性低的项目应删除，因此需要对题库进行主成分分析(PCA)，删除在第一主成分载荷小于 0.4 的项目。

接着对保留的项目进行单维性检验，单维性是项目反应理论的前提假设之一(Hambleton et al., 1991)，而已有研究表明，在探索性因素分析(EFA)中，第一特征值与第二特征值比值大于 4 且第一因子解释方差大于 20%，则可以认为项目满足单维性假设(Reckase, 1979; Andrich, 1996; Reeve et al., 2007)。

2.3.2IRT 模型选择

因本研究的项目均为多级计分，可选的 IRT 模型主要有拓广分部评分模型(GPCM；Muraki, 1992)与等级反应模型(GRM；Samejima, 1969)，因此本研究将比较两个模型的拟合指数，主要是 AIC(Akaike, 1974)与 BIC(Schwarz, 1978)，选择拟合更优的模型进行后续的参数估计。

2.3.3 项目分析

词嵌入(word embedding)是 NLP 模型的重要部分，其表现出的一致、普遍的性别偏见引发了研究者的关注(Gonen & Goldberg, 2019; Lee et al., 2023)。词嵌入指将每个单词表示成一定维度的向量，在词向量空间中，不同词之间的关系可以用这些词向量的差异来捕获(Bolukbasi et al., 2016; Caliskan & Lewis, 2020)。现实中的一些刻板印象，就可能导致词嵌入技术错误地捕获了这些偏见(Garg et al., 2018)。为避免性别偏见对 ES 测量引起的系统差异，我们将进行项目功能差异检验(Differential item functioning，DIF)。当来自不同群体的受测者在匹配项目欲测量的潜在特质水平后，仍在该项目上表现出不同的统计特性，那么就说明这个项目存在了 DIF(Zumbo, 1999)。本研究采用逻辑回归的方法来检验是否存在性别引

起的 DIF，当 McFadden's R^2 大于 0.02 时，表明该项目存在 DIF，需要考虑删除(Choi et al., 2011)。其计算公式如下：

$$\text{McFadden's } R^2=1-\frac{\ln L}{\ln L_0}$$

$$\text{模型 1: logit } P(u_i \geqslant k)=\alpha_k+\beta_1*\theta$$

$$\text{模型 2: logit } P(u_i \geqslant k)=\alpha_k+\beta_1*\theta+\beta_2*\text{性别}$$

$$\text{模型 3: logit } P(u_i \geqslant k)=\alpha_k+\beta_1*\theta+\beta_2*\text{性别}+\beta_3*\theta*\text{性别}$$

其中 L_0 代表基线模型的似然值，L 代表增加预测变量后模型的似然值。$P(u_i \geqslant k)$代表对于第 i 个项目，作答为第 k 个及更高水平的选项的累积概率（1≤k≤项目总选项数），α 代表回归模型的截距，β 则代表回归系数，θ 代表项目欲测量的特质水平（本研究中即为情绪稳定性）。当检验一致性 DIF 时，模型 1 便是基线模型，模型 2 为增加预测变量后的模型；检验非一致性 DIF 时，模型 2 是基线模型，而模型 3 则为增加预测变量后的模型。

另外项目区分度也是评估项目质量的重要指标：区分度指项目对不同能力水平的被试的鉴别力，其数值越大代表能更好区分不同能力水平的被试。

2.3.4 题库信息量与边际信度

项目信息量代表项目在评价被试特质水平时提供信息的确定性水平，其值越大，表明项目的可靠性越高。而测验信息量是所有测验项目信息量之和，其值的平方与测验标准误成反比。计算公式如下：

$$SE(\theta)=\sqrt{\frac{1}{\sum_{i=1}^{m} I_i(\theta)}}$$

其中 θ 代表被试的潜在特质水平，本研究中即为情绪稳定性水平，m 为测验的项目总数，$I_i(\theta)$表明第 i 个项目对于特质水平为 θ 的受测者提供的信息量。

整个测验整体的可靠性有研究者使用边际信度(MR)来表示(Liu, 2022; Xu et al., 2020)，其通过所有受测者的平均测量标准误计算得到。公式如下：

$$SE(\theta)=\frac{\sum_{i=1}^{m} SE(\theta_i)}{N} \qquad MR=1-SE^2$$

其中 N 代表受测者总人数，i 代表第 i 个受测者，$SE(\theta_i)$代表第 i 个受测者在最终估计 θ 时的测量标准误。

2.3.5 模拟 CAT

在得到最终确立的题库后，根据 479 位受测者在全部项目上的真实作答情况分别进行基于三个题库的模拟 CAT。CAT 中使用的选题策略采用目前最广泛使用的最大信息量法

(MFI)，能力估计方法采用期望后验估计法(EAP)。终止规则方面，首先采用答完总题库中所有项目的规则，将该条件下得到的能力值视为能力真值，然后我们采用固定测量精度的规则，即当能力估计的标准误达到特定值时才停止测验（本研究采用边际信度=0.8/0.85/0.90/0.95 对应的 SE =0.447/0.387/0.316/0.224，以及对四个已有测验实际测量对应的 SE= 0.34/0.27/0.46/0.21），通过比较 3 个题库在不同终止条件下所需的项目数量、能力估计的标准误以及与能力真值的相关系数等，来考察其 CAT 性能；接着，我们采用定长的终止规则，即作答的项目数量达到规定值后便停止测验（长度对应四个已有量表中测量 ES 项目的数量=12/8/20/2），同时估计出被试以传统测验形式分别作答 4 个量表的 ES 项目时的能力值与测量误差，比较 3 个题库 CAT 与传统测验形式下的测量误差、边际信度以及能力的相关系数等，接着再与样本 2、3 在传统测验上的结果进行信度比较，多方面探索 GPT 题库进行 CAT 的可行性，以及 CAT 相较于传统测验形式的性能提升。

效度是 CAT 性能另一重要指标(Xu et al., 2020)，本研究将效标效度作为参考指标，只有当 CAT 的评估结果与校标量表的测量结果相似时，才能认为 CAT 是有效的。本研究将 CBF-PI-B、BFI-2、TIPI、BFAS 中的 ES 维度作为校标，分别计算被试在三个题库中作答所有项目后的能力估计值与这 4 个量表上得分的相关系数，以此来验证 3 个题库的效度。

2.4 研究工具

本研究采用基于 OpenAI 于 2023 年 11 月初发布的 GPT-4 Turbo 版本的 ChatGPT 进行项目生成，接着采用 SPSS 26.0 进行主成分分析和单维性检验，其余分析则采用 R 软件包，如使用 mirt 包进行 IRT 模型选择、项目区分度分析、题库信息量计算等，lordif 包检验项目功能差异，最后 catR 包进行模拟 CAT。

3 研究结果

3.1 题库构建

3.1.1 单维性检验

分别对两个题库的项目进行主成分分析(PCA)，结果表明所有项目在第一主成分上的载荷量均大于 0.4（详情可见图 1），因此所有项目均得以保留。

接着对经典题库、GPT 题库分别进行了 KMO 测试，其结果为 0.979、0.986，证实了数据进行因子分析的适用性(χ^2 (861) = 17662.03, $p < 0.001$；χ^2 (2775) = 35225.09, $p < 0.001$)。之后便采用主轴因式分解(PAF)以及最优斜交法进行了 EFA，结果显示所有因子载荷均大于 0.4，经典题库的第一、第二特征值分别为 22.75，2.29，比值为 9.92，且第一因子解释方差

达到 53.28%，GPT 题库的第一、第二特征值分别为 42.93，1.97，比值为 21.78，第一因子解释方差达到 56.79%，因此表明两个题库中的项目均满足单维性假设。

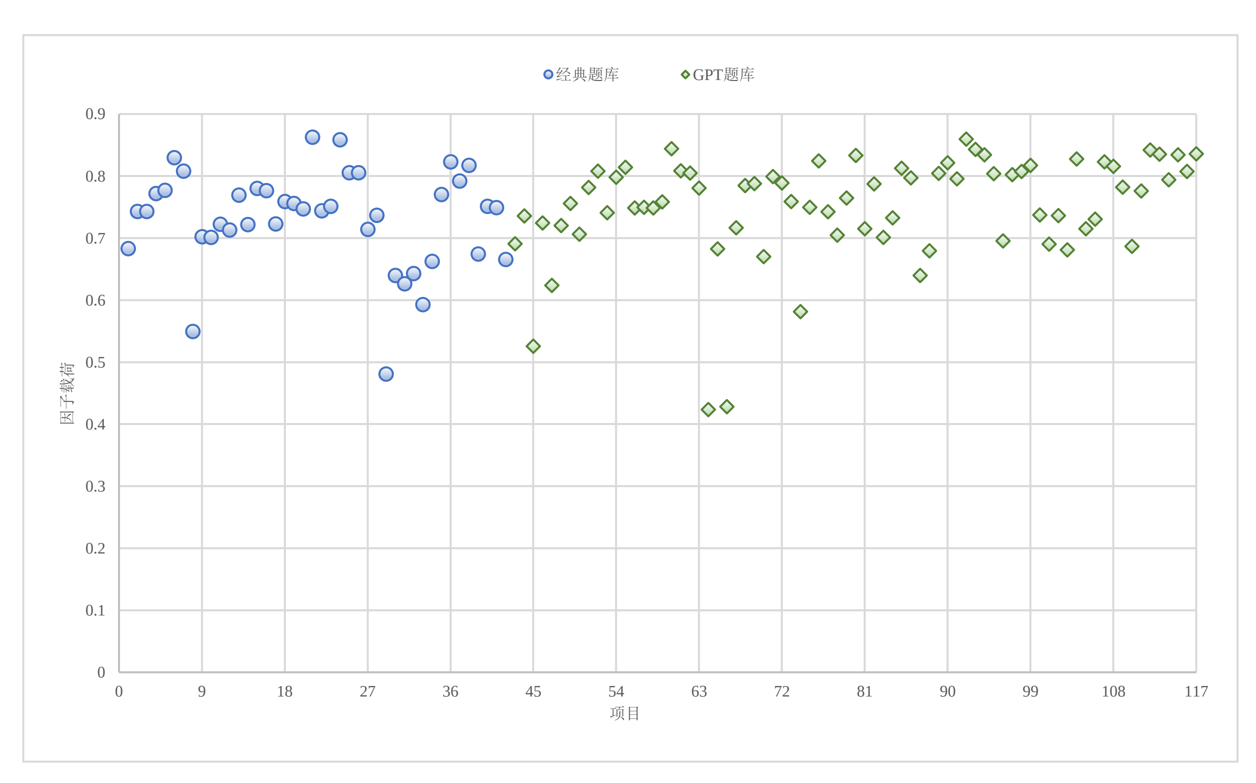


图 1 两个题库项目的因子载荷

3.1.2 项目反应理论模型选择

结果如表 1 所示，两个题库在 GRM 拟合中的 AIC 与 BIC 均更小，表明其拟合效果更优，因此 GRM 被用于后续的分析。

表 1 模型拟合指标值

模型	AIC	BIC	Loglik
经典题库			
GRM	47240.97	48167.08	–23398.48
GPCM	47820.24	48746.35	–23688.12
GPT 题库			
GRM	74640.70	76242.64	–36936.35
GPCM	75556.12	77158.05	–37394.06

3.1.3 项目分析

项目功能差异检验结果显示所有项目 McFadden's R^2 值均小于 0.02，表明项目不存在性别引起的项目功能差异，因而均得以保留。

一般认为区分度大于 0.8 的项目较优（Liu, 2022），因此低于该标准的项目需要删除。分别对两个题库再次拟合 GRM 模型，结果发现两个题库中所有项目的区分度均大于 0.8

（详情可见图 2），均值分别为 2.27(*SD*=0.51)、2.44(*SD*=0.58)。难度区间则分别为[－3.1, 2.2]与[－4.3, 2.1]，表明两个题库难度覆盖范围较广。总的来说，两个题库质量均较高，GPT 生成项目的区分度总体上甚至略优于已有的项目，并有更广的难度分布。

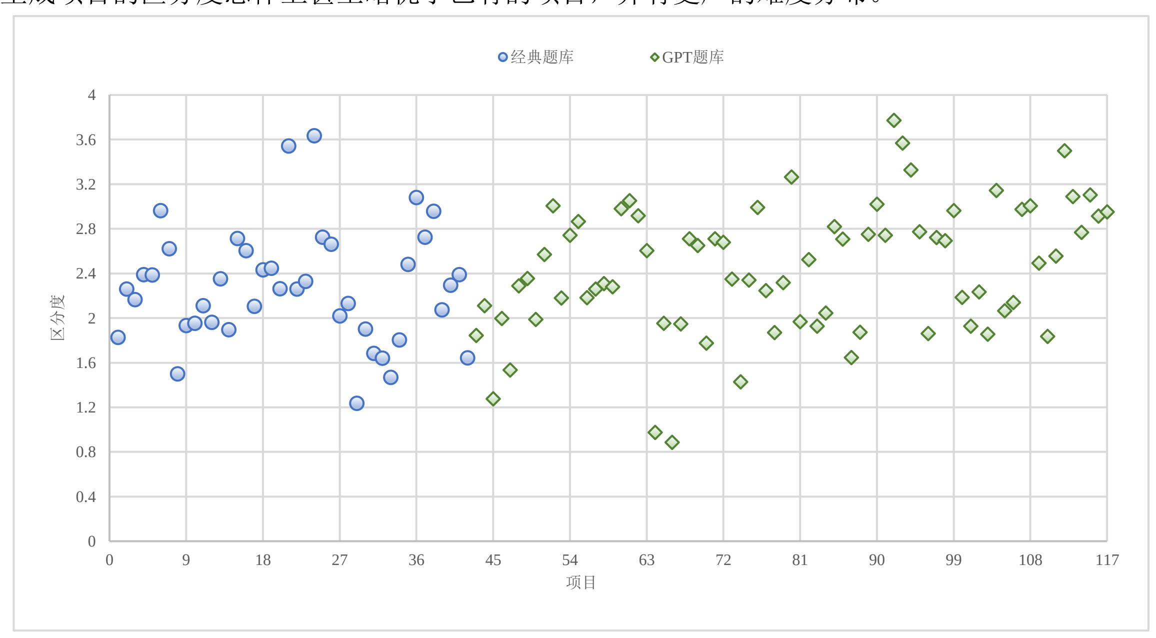


图 2 两个题库中项目区分度

3. 1. 4 题库信息量与边际信度

本研究中两个题库的测验信息量及其标准误如图 3 所示，一般认为不高于 0.39 是低测量标准误的界限(Xu et al., 2020)，总的来说两个题库均能为大部分受测者提供较高的信息量及较低的测量误差，仅对于情绪稳定性水平极高的被试有一定局限。此外，两个题库的边际信度如图 4 所示，其平均值分别高达 0.96，0.98，表明两个题库整体可靠性均很高。相较而言，GPT 题库整体质量较经典题库更高，尤其对于特质水平较高的受测者，在 GPT 题库的测量标准误仍处于低标准内，并有更高的信度。

3. 1. 5 总题库构建

将两个题库中最终保留的共计 117 道题组成总题库。对总题库的项目进行主成分分析，同样得到所有项目在第一主成分载荷量均大于 0.4（详情见附录图 6）。KMO 测试结果为 0.986(χ^2 (6786) = 57274.54, p < 0.001)。接着采用与前文相同的方法进行了 EFA，结果仍显示所有因子载荷均大于 0.4，第一、第二特征值分别为 62.98，4.47，比值为 14.10，第一因子解释方差达到 53.83%，因此表明总题库中 117 个项目满足单维性假设。

拟合指标结果同样表明 GRM 更优（详情见附录表 8），因此再次使用该模型估计项目参数，结果表明所有项目区分度仍均高于 0.8，均值为 2.25(*SD*=0.50)（详情见附录图 7）。最后对总题库的信息量、测量标准误以及边际信度进行计算（详情见附录图 8、图 9），总

体上看总题库相较于经典题库、GPT 题库有更小的测量误差与更高的可靠性，即使对于情绪稳定性高的个体也能提供较为精确的测量结果，其平均边际信度更是高达 0.99。

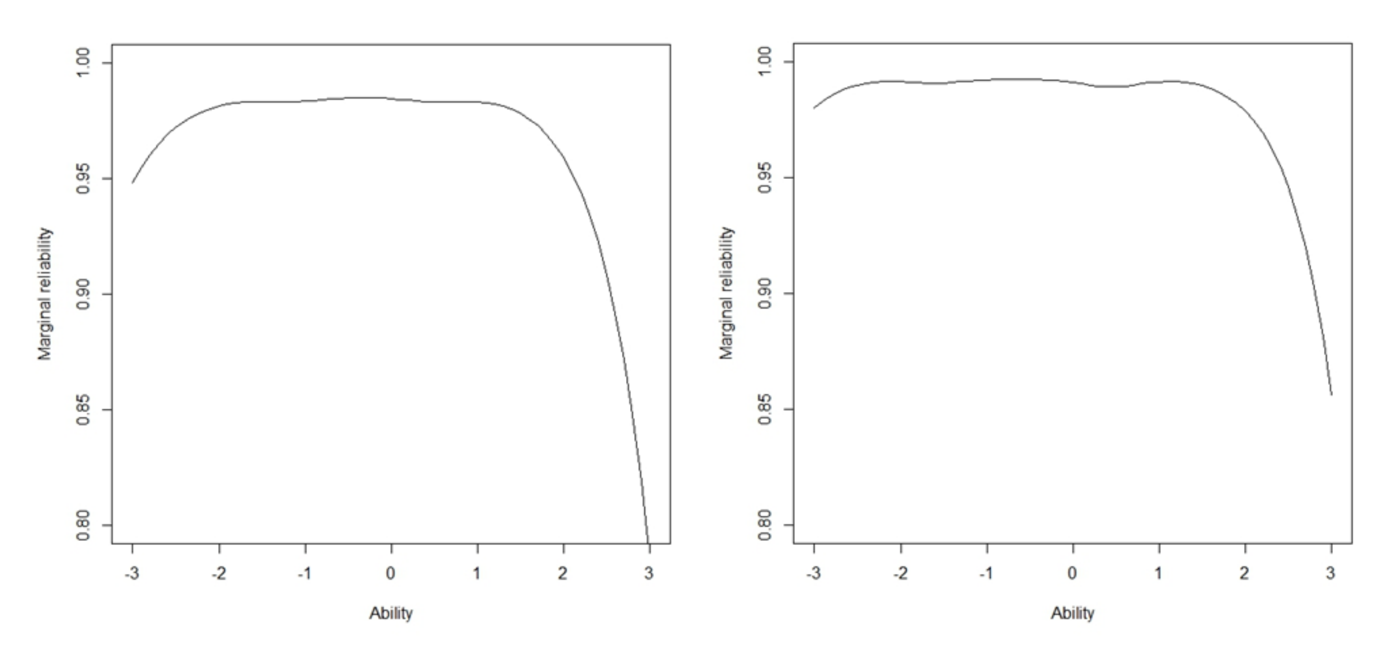


图 3 经典题库（左）与 GPT 题库（右）的信息量及标准误

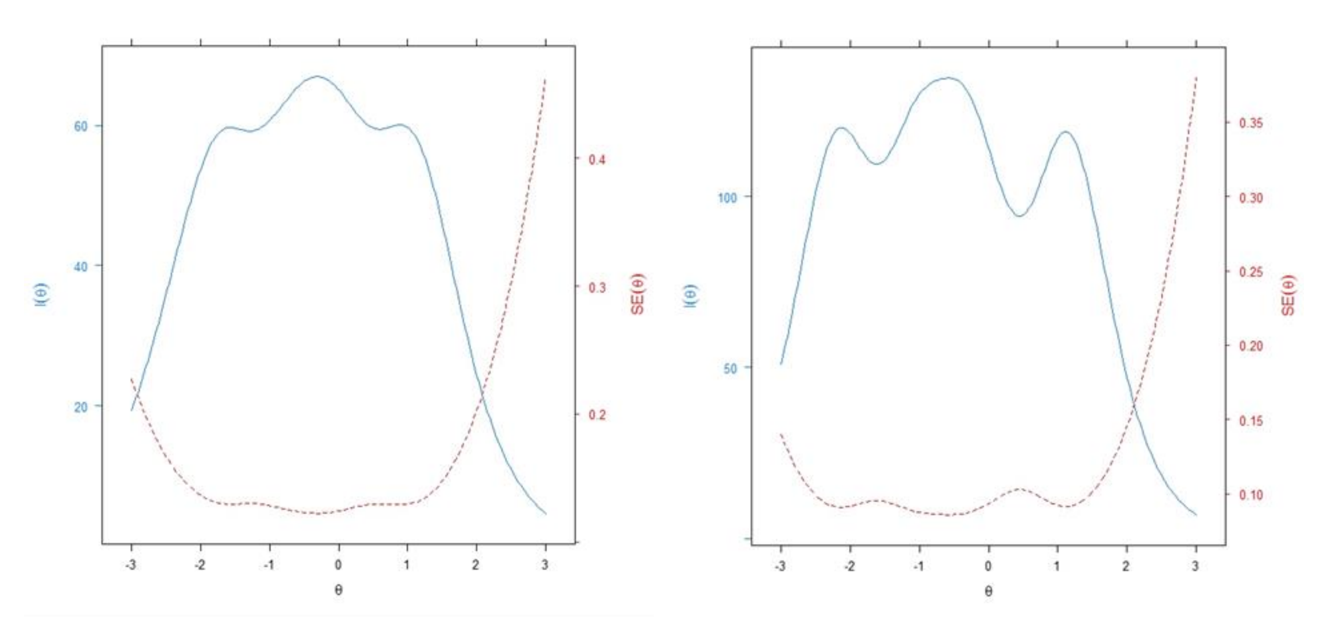


图 4 经典题库（左）与 GPT 题库（右）的边际信度曲线

3.2 模拟 CAT

3.2.1 定测验精度条件下 3 个题库模拟 CAT 的表现

表 2 显示了 3 个题库在不同固定精度的停止规则下模拟 CAT 的结果。可以发现，即使在 SE(θ)≤0.447 的情况下，在 3 个题库中均仅平均使用不到三个项目便能使能力估计值与真

值的相关系数高达 0.9 以上(n=479, p＜0.001)，但此时边际信度不够理想。一般认为信度系数达到 0.85 表明测验具有较高的可靠性（张龙飞等，2020; May et al., 2006）。而 3 个题库达到 0.87 的边际信度仅需要使用 3.06-3.77 个项目，即使 0.91 的信度也仅需要 4.79-5.87 个项目，如果要求更高的可靠性，则作答项目数将是之前的一倍以上，需要 10.11-12.90 题，但此时边际信度可以达到 0.95 且与能力真值的相关系数达到 0.97 以上(n=479, p＜0.001)。为进一步直观体现 CAT 的效率，表 3 对比了三个题库在 CAT 方式下，达到传统测验的测量误差下所需题数，总体上节省了 24.4%-47.5%的项目，仅相较 TIPI 的 2 道项目略有增加，这可能是由于 CAT 在测试初始阶段由于无法对受测者的能力进行估计而只能在题库中随机选择项目造成的。

通过比较经典题库、GPT 题库的结果，可以发现在相同停止精度下， GPT 制定的题库所需的测验题数均少于由经典项目组成的题库，且其边际信度及与能力真值的相关系数也与经典题库的结果相似甚至略优。同时，达到传统测验精度的情况下，GPT 题库所需项目数显著低于经典题库，以达到 BFAS 量表的精度为例，GPT 题库相较经典题库显著减少了 23.13%的项目数($t_{(478)}$=17.20, p<0.001, Cohen's d=0.79)。而总题库大体上与 GPT 题库的 CAT 的性能相似。

综上，在一定测量精度的条件下， GPT 题库的 CAT 性能甚至优于经典题库，同时 CAT 的形式相较于传统测验提升了测验效率，在保证测量准确的情况下进一步减少了测验所需的项目数量。

3.2.2 定测验长度条件下 3 个题库模拟 CAT 的表现

表 4、表 5 分别展示了当以传统测验的题数为 CAT 的终止条件时，3 个题库 CAT 相较于传统测验的测量误差降低与边际信度提高的情况。基于经典题库的 CAT 较除 TIPI 外的传统测验显著降低了 13.83%-22.69%的测验误差(p<0.001, Cohen's d>1)，显著提升了 1.28%-5.40%的测量信度(p<0.001, Cohen's d>1)，仅较 TIPI 没有显著差异(p>0.05)。而 GPT 题库、总题库的 CAT 均较 4 个传统测验有显著的性能提升，分别降低了 6.50%-31.44%与 7.91%-31.65%的测量误差(p<0.001)，提高了 2.20%-7.14%与 2.30%-7.17%的测验信度(p<0.001)。表 6 则展示了 3 个题库 CAT 较不同样本中两个传统测验的性能提升情况，可以发现总体上与在样本 1 中的比较结果相近，即使相对于更大样本量的传统测验，3 个题库的 CAT 均有显著的信度提高（p<0.001, Cohen's d>1）。

表 2 不同精度停止条件下的结果

题库	终止规则	使用的项目数量		*Mean SE*(θ)	边际信度	能力相关系数
		M	*SD*			
经典题库						
	全部	42	0.00	0.15	0.98	.986***
	SE(θ) ≤0.447	2.60	0.88	0.41	0.83	.900***
	SE(θ) ≤0.387	3.77	1.73	0.36	0.87	.929***
	SE(θ) ≤0.316	5.87	3.23	0.30	0.91	.950***
	SE(θ) ≤0.224	12.90	4.77	0.22	0.95	.978***
GPT 题库						
	全部	75	0.00	0.09	0.99	.992***
	SE(θ) ≤0.447	2.40	1.06	0.40	0.84	.921***
	SE(θ) ≤0.387	3.06	1.18	0.36	0.87	.930***
	SE(θ) ≤0.316	4.90	3.76	0.30	0.91	.953***
	SE(θ) ≤0.224	10.11	6.94	0.22	0.95	.976***
总题库						
	全部	117	0.00	0.08	0.99	1.00
	SE(θ) ≤0.447	2.42	0.96	0.39	0.84	.916***
	SE(θ) ≤0.387	3.15	1.39	0.36	0.87	.932***
	SE(θ) ≤0.316	4.79	3.30	0.30	0.91	.948***
	SE(θ) ≤0.224	10.14	8.38	0.22	0.95	.971***

注：***代表 p<0.001，下同

表 3 达到传统测验精度所需项目数

测量误差	传统测验长度	经典题库长度	GPT 题库长度	总题库长度
SE(CBF-PI-B) =0.34	8	5.07	4.18	4.20
SE(BFI-2) =0.27	12	8.44	6.86	7.08
SE(TIPI) =0.46	2	2.51	2.27	2.27
SE(BFAS) =0.21	20	15.13	11.63	11.52

此外，通过比较可以发现，GPT 题库相较于经典题库，在同样的测验长度下，显著降低了测量误差($t_{(478)}$= 26.38、30.35、12.07、37.13, p<0.001)并显著提高了测量信度($t_{(478)}$= 22.53、24.68、11.13、29.28, p<0.001)，尤其是在测验长度等于 TIPI 的 2 题的情况下，GPT 题库将经典题库相较于传统测验的负提升转化为了正提升。而总题库总体性能略优于 GPT 题库。

最后我们比较了在相同测验长度条件下，传统测验方法以及 3 个题库 CAT 方法下能力估计值间的相关系数以及与能力真值的相关系数并以热力图的形式进行展示，来直观考察 CAT 测量的准确性与稳定性。从图 5 可以看出，3 个题库在不同测验长度下，与传统测验的结果相关性从 0.79-0.97 不等（均值=0.90）。且 3 个题库下能力估计值与能力真值的相关性均比相同长度下传统测验能力估计值与能力真值的相关更高，以上均表明 CAT 测验形式准确性较高。另外，我们不难发现，GPT 题库得到的能力估计值与能力真值的相关系数普遍

高于经典题库，由此证明了 GPT 题库相较于经典项目组成的题库在准确测量上体现出的优越性。

表 4 相同测验长度时不同方法的测量误差

测验方法	测验长度	平均能力值 (*SD*)	测量误差			
			M(SD)	*SE* 降低百分比	*t*	Cohen's *d*
CBF-PI-B	8	0.08 (0.87)	0.34 (0.03)			
经典题库	8	0.03 (0.98)	0.27 (0.03)	22.69%	48.06***	2.20
GPT 题库	8	0.06 (0.98)	0.24 (0.03)	31.44%	72.67***	3.32
总题库	8	0.06 (0.99)	0.23 (0.03)	31.65%	68.15***	3.11
BFI2	12	0.06 (0.98)	0.27 (0.02)			
经典题库	12	0.05 (0.99)	0.23 (0.03)	15.24%	43.73***	2.00
GPT 题库	12	0.06 (1.00)	0.20 (0.03)	25.51%	65.97***	3.01
总题库	12	0.07 (0.98)	0.19 (0.03)	26.65%	70.91***	3.24
TIPI	2	0.08 (0.87)	0.46 (0.05)			
经典题库	2	0.06 (0.92)	0.46 (0.05)	–0.81%	–1.27	–0.06
GPT 题库	2	0.03 (0.94)	0.43 (0.05)	6.50%	10.78***	0.49
总题库	2	0.08 (0.89)	0.42 (0.05)	7.91%	13.43***	0.61
BFAS	20	0.05 (1.00)	0.21 (0.03)			
经典题库	20	0.06 (0.99)	0.19 (0.02)	13.83%	41.10***	1.88
GPT 题库	20	0.06 (1.01)	0.16 (0.03)	26.06%	72.31***	3.30
总题库	20	0.06 (1.01)	0.16 (0.02)	27.01%	78.38***	3.58

表 5 相同测验长度时不同方法的测量信度

测验方法	测验长度	测量信度			
		M(SD)	*MR* 增加百分比	*t*	Cohen's *d*
CBF-PI-B	8	0.88 (0.03)			
经典题库	8	0.93 (0.02)	5.40%	41.97***	1.92
GPT 题库	8	0.94 (0.02)	7.14%	61.99***	2.83
总题库	8	0.94 (0.02)	7.17%	57.68***	2.64
BFI-2	12	0.93 (0.02)			
经典题库	12	0.95 (0.01)	2.14%	39.20***	1.79
GPT 题库	12	0.96 (0.02)	3.35%	57.61***	2.63
总题库	12	0.96 (0.01)	3.50%	61.47***	2.81
TIPI	2	0.79 (0.05)			
经典题库	2	0.78 (0.05)	–0.47%	–1.29	–0.06
GPT 题库	2	0.81 (0.05)	3.27%	9.87***	0.45
总题库	2	0.82 (0.05)	3.98%	12.34***	0.56
BFAS	20	0.95 (0.02)			
经典题库	20	0.97 (0.01)	1.28%	29.91***	1.37
GPT 题库	20	0.97 (0.01)	2.20%	51.91***	2.37
总题库	20	0.97 (0.01)	2.30%	48.12***	2.20

表 6 跨样本下相同测验长度时的测量信度

测验方法	测验长度	测量信度			
		M(SD)	*MR* 增加百分比	*t*	Cohen's *d*
样本 2 CBF-PI-B	8	0.88 (0.03)			
经典题库	8	0.93 (0.02)	5.39%	37.48***	1.87
GPT 题库	8	0.94 (0.02)	7.12%	50.83***	2.54
总题库	8	0.94 (0.02)	7.13%	50.66***	2.53
样本 3 BFI-2	12	0.93 (0.01)			
经典题库	12	0.95 (0.01)	1.77%	20.16***	1.21
GPT 题库	12	0.96 (0.02)	2.98%	31.23***	1.88
总题库	12	0.96 (0.01)	3.13%	36.58***	2.20

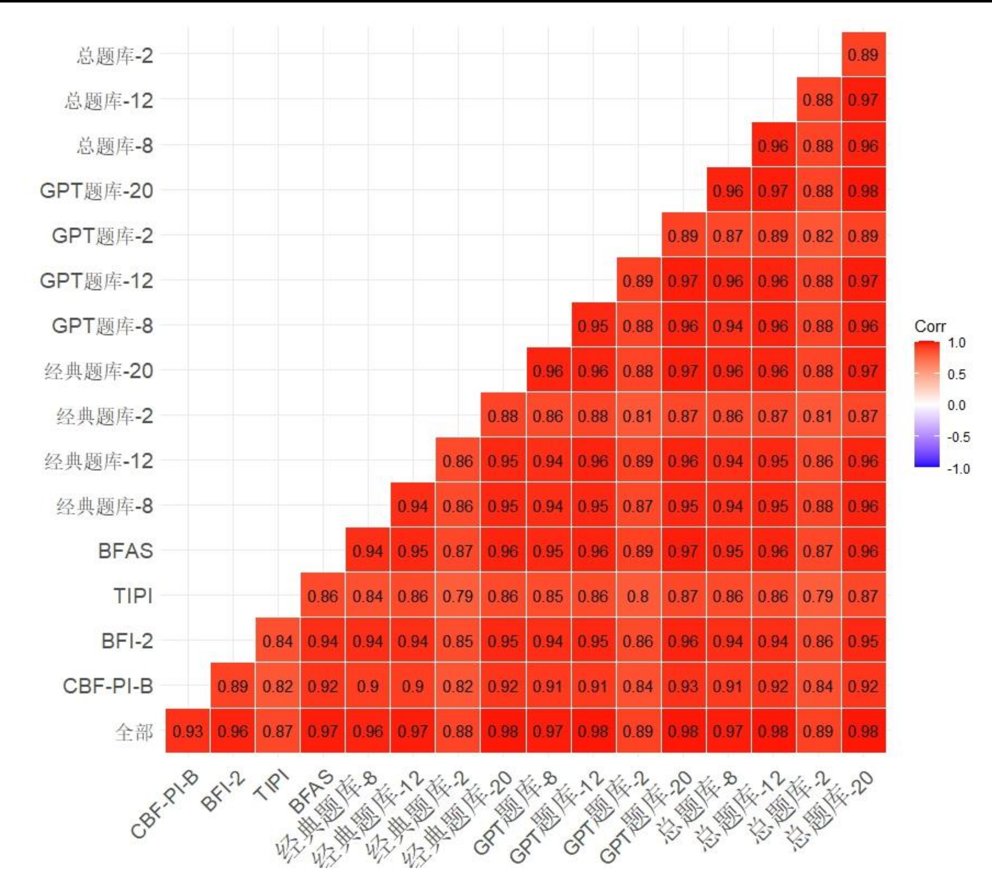


图 5 不同定长终止规则下相关系数的热力图

3.2.3 效度验证

效度验证结果如表 7 所示，从中不难发现 3 个题库均与四个校标量表得分存在显著的相关($p<0.001$)且相关系数均大于 0.83，说明 3 个题库具有理想的效度。

表 7 3 个题库校标关联效度

测量误差	CBF-PI-B	BFI-2	TIPI	BFAS
经典题库	0.832***	0.913***	0.860***	0.914***
GPT 题库	0.831***	0.917***	0.861***	0.918***
总题库	0.836***	0.921***	0.862***	0.919***

4 讨论

本研究旨在利用最新版本的 ChatGPT 生成大量中文版测量情绪稳定性的人格项目，并探索这些项目构建的题库对于 CAT 的适用性。通过结果可以看出，GPT 题库有良好的质量，并展现出较经典项目更优异的 CAT 性能。最终的总题库则在有足够效度的同时，相较于传统测验在测量精度与效率方面都有较大的提升。

本研究创新性地运用 GPT 构建中文版计算机自适应题库。传统的项目编写需要专家利用经验与知识构建新项目，并不断审查、修改与完善，直至满足所需的质量标准，这一迭代过程既耗时又昂贵(Gierl et al., 2012)，而本研究的方法为构建题库提供了一种高效且经济的途径。同时 ChatGPT 这一对话的使用形式，相较于前人所使用的旧版本(Hernandez & Nie, 2022; Götz et al., 2023; Lee et al., 2023)，不再要求研究者拥有扎实的 NLP 与机器学习基础以针对项目开发任务进行模型微调，甚至不再需要进行编程，极大降低了项目开发者的上手门槛。

基于 GPT-4 的 ChatGPT 在保证了简易使用形式的同时，产生了更优质的项目。通过与前人研究结果对比发现，本研究生成的项目有更高的效度，例如同样是将 GPT 生成的项目与 BFI-2 相关，在情绪稳定性这一维度下，GPT-4 的相关系数 0.917（p<0.001）远高于前人 GPT-2 的 0.786 (p<0.001) (Götz et al., 2023) 。另一方面，GPT-4 产生的无效项目更少，经过人工质量评判后，本研究项目保留率达到了 65.78%，而 Götz 等人 (2023)通过 10 次迭代生成的 1 万个项目中，有 60%与示例项目完全重复，经专家审查后最终仅保留了 92 个项目。同样地， Hommel 等人（2022）生成的 1360 个项目中，完全与示例项目重复的便有 1077 个，剩余项目也仅有 53.4%获得了专家对内容有效性的认可。从成千上万的项目中进行筛选，无疑既繁琐又与利用 AIG 技术节省时间与人力成本的初衷相悖，而 NLP 技术的进步很大程度上解决了这一痛点。

通过本研究的探索可以发现 GPT 在生成非认知项目方面有巨大潜力，自动项目生成逐渐展现出的对于人工项目编制的替代性，也提示着心理学人的职责转变。由于使用 GPT 按照特定测量目标生成项目的具体过程是复杂且模糊的，使得该类方法有一定的不可预测性(Hommel et al., 2022)，因此后续项目生成过程中需要研究者更关注对测量目标进行准确详细的定义，以使算法尽可能生成结构效度高的项目，正如 Bejar(2013)提出的那样“项目生成和概念表征是相辅相成的”(p.43)。此外，不仅项目质量仍需要专家制定评判标准并对生

成的项目进行审核，如项目的内容是否符合测量目标、项目是否存在语法或用词上的错误、项目是否存在偏见等等，而且标准化项目生成的流程、为 GPT 制定全面且完善的项目生成原则、准确定义与描述测量目标等都是后续研究者需要努力的方向。通过最近 OpenAI 推出的允许用户根据自定义规则创建专用版本 ChatGPT 的 GPTs 功能，可见在通用大语言模型的基础上，专业学者利用知识与经验定制个性化、专业化模型来执行特定任务是必然趋势。

尽管本研究为基于最新 NLP 技术的 AIG 与 CAT 结合提供了有力证据，但仍然存在一些局限之处：一方面，GPT 生成的项目仍存在一些不足，例如缺乏反向计分的项目、项目质量参差不齐等；另一方面，CAT 的相关算法是否有更优的选择也有待商榷，例如基于 KL 全局信息量(Chang & Ying, 1996)的选题法相较于 Fisher 信息量可能是更稳健的选择，后续还可以采用一些约束条件控制项目的曝光率（如最大优先级指标法，Cheng & Chang, 2009）。

总的来说，NLP 技术的发展为非认知项目的 AIG 及以此为基础的 CAT 提供了易上手且表现良好的工具，本研究利用了最新版本的 ChatGPT，展示了大语言模型在计算机自适应题库生成中的巨大潜力与可能性。在未来的研究中，GPT 能否应用于更广泛测量目标的项目生成，尤其是针对一些缺乏已有测量工具的目标概念，需要进一步探索，同时除了经典的李克特式评级项目，迫选题(A. Brown & Maydeu-Olivares, 2011)等更丰富的测验形式能否借助于 GPT 得到进一步的发展也令人期待。

附录

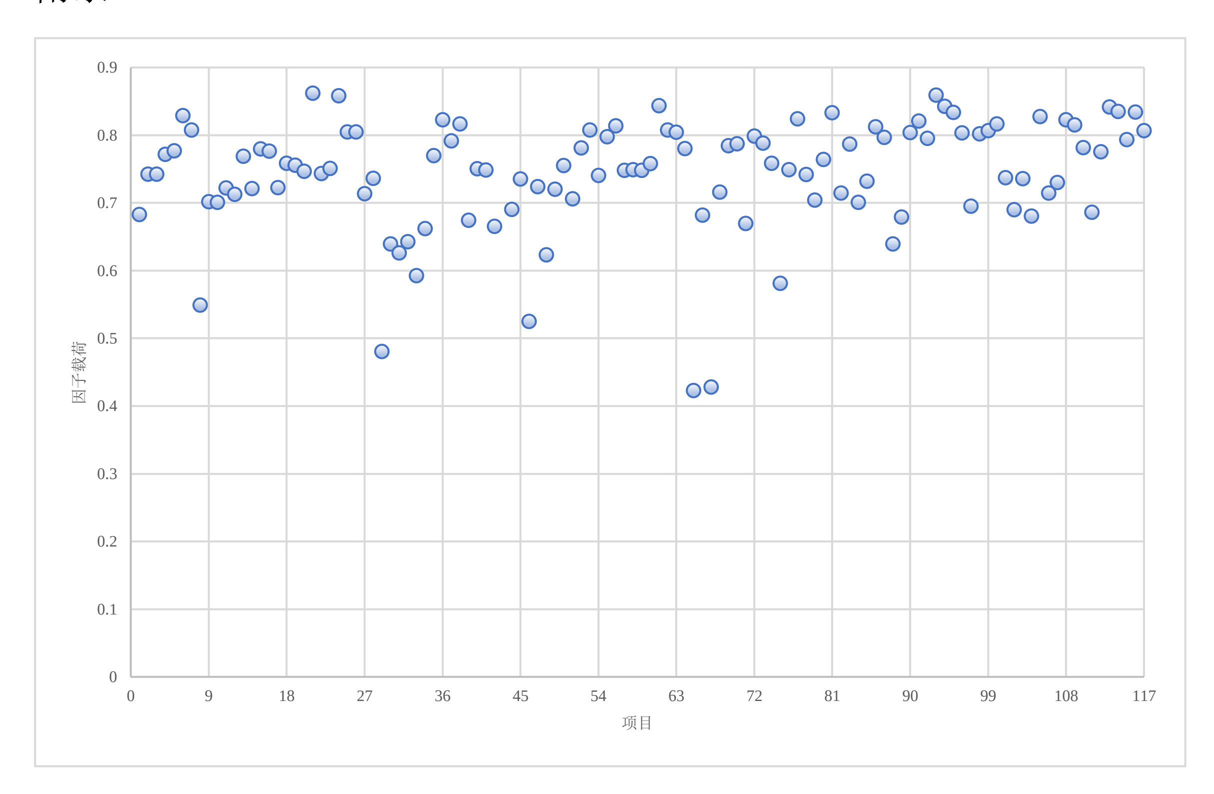


图 6 总题库所有项目的因子载荷

表 8 总题库模型拟合指标值

模型	AIC	BIC	Loglik
GPCM	124201.5	126729.6	–61494.76
GRM	122730.8	125258.8	–60759.38

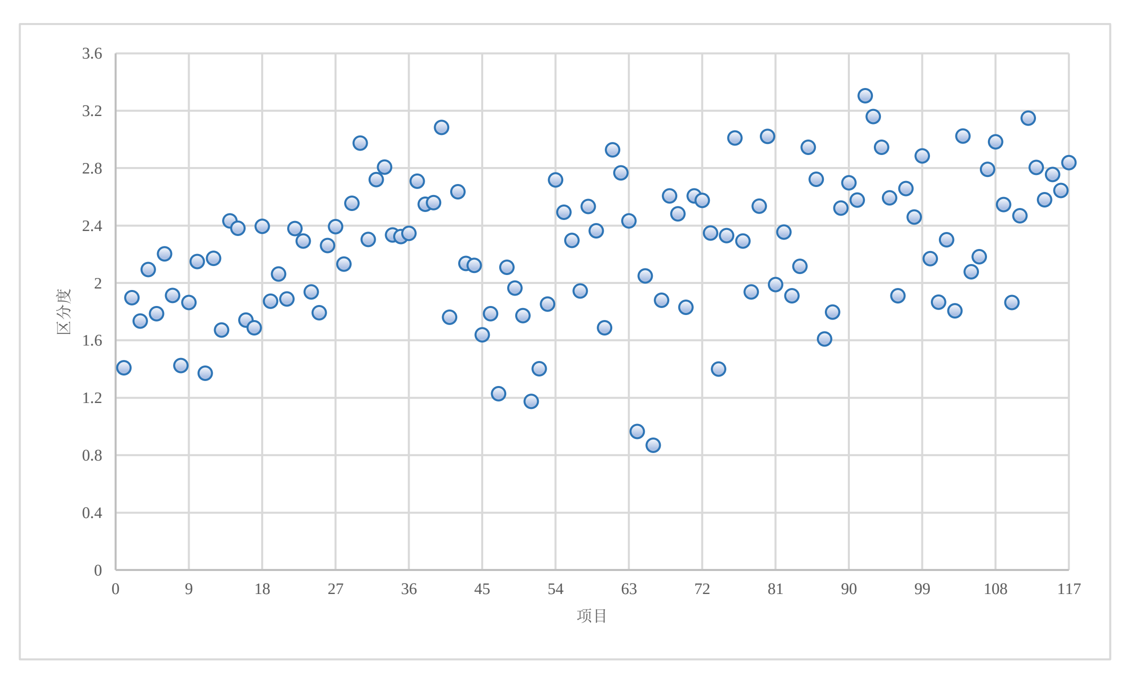


图 7 总题库中所有项目的区分度

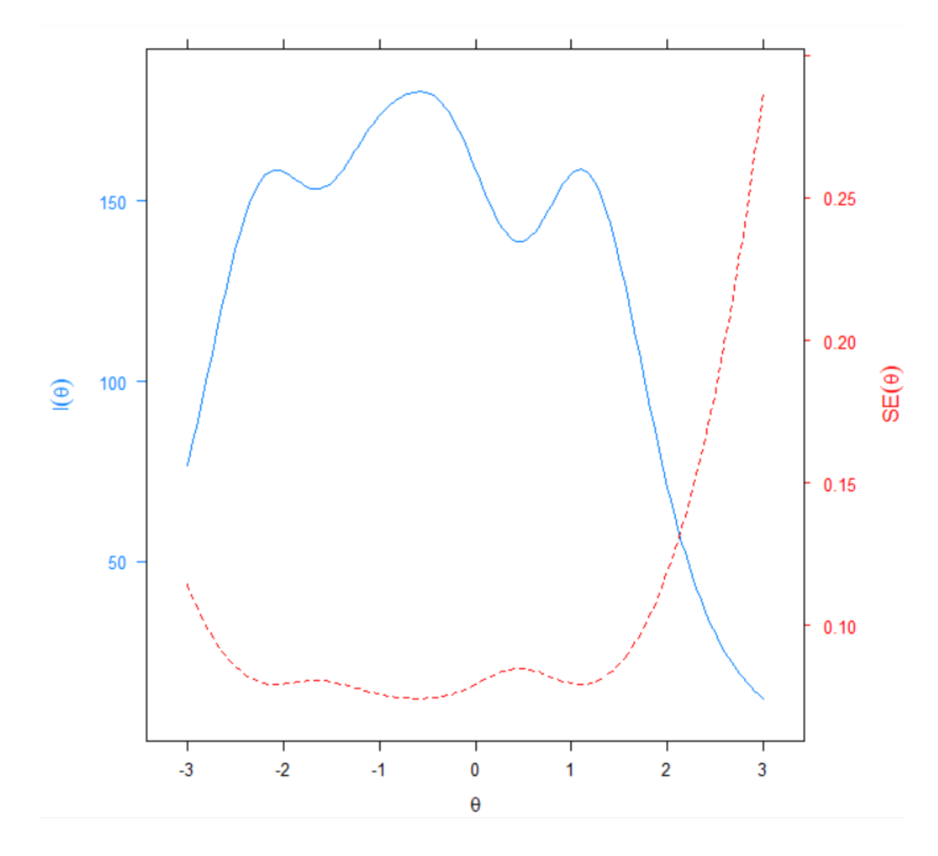


图 8 总题库信息量及标准误

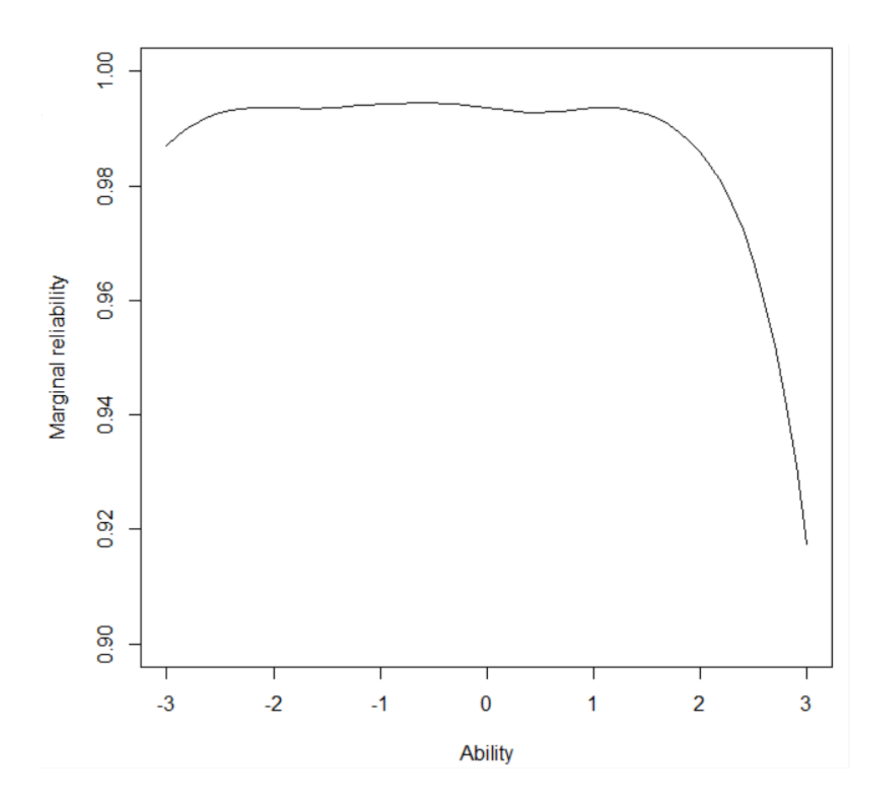


图 9 总题库边际信度曲线

参考文献

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akaike, H. (1974). Stochastic theory of minimal realization. *IEEE Trans. Automat. Control* 19, 667–674. doi: 10.1109/cdc.1976.267680

Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling thurst-one and likert methodologies. *Br. J. Math. Stat. Psychol*. 49, 347–365. doi: 10.1177/014662169301700307

Bajaj, B., Gupta, R., & Sengupta, S. (2019). Emotional Stability and Self-Esteem as Mediators Between Mindfulness and Happiness. *Journal of Happiness Studies, 20*(7), 2211–2226. https://doi.org/10.1007/s10902-018-0046-4

Bec, A., & Becken, S. (2021). Risk perceptions and emotional stability in response to Cyclone Debbie: An analysis of Twitter data. *Journal of Risk Research*, *24*(6), 721–739. https://doi.org/10.1080/13669877.2019.1673798

Bejar, I. (2013). Item generation: Implications for a validity argument. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 40–55). Routledge.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Brown, A., & Maydeu-Olivares, A. (2011). Item Response Modeling of Forced-Choice Questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

Caliskan, A., & Lewis, M. (2020). *Social biases in word embeddings and their relation to human cognition* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/d84kg

Chang, H.-H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement*, *20*(3), 213–229. https://doi.org/10.1177/014662169602000303

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 369–383. https://doi.org/10.1348/000711008X304376

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). **lordif**: An *R* Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations.

Journal of Statistical Software, *39*(8). https://doi.org/10.18637/jss.v039.i08

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a Computer-adaptive Test for Depression (D-CAT). *Quality of Life Research*, *14*(10), 2277–2291. https://doi.org/10.1007/s11136-005-6651-9

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16). https://doi.org/10.1073/pnas.1720347115

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M. J., & Lai, H. (2018). Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing. *Applied Psychological Measurement*, *42*(1), 42–57. https://doi.org/10.1177/0146621617726788

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items: Automatic generation of test items. *Medical Education*, *46*(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies,* -, 1-309.

Gonen, H., & Goldberg, Y. (2019). *Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them* (arXiv:1903.03862). arXiv. http://arxiv.org/abs/1903.03862

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Götz, F. M., Maertens, R., Loomba, S., & Van Der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*. https://doi.org/10.1037/met0000540

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications Inc. doi: 10.2307/2075521

Hernandez, I., & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, peps.12543. https://doi.org/10.1111/peps.12543

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-Based Deep

Neural Language Modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, *87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Jiao, H., & Lissitz, R. W. (Eds.). (2020). *Application of artificial intelligence to assessment*. Information Age Publishing, inc.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, *3(2)*, 114-158.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A.-P., & De Champlain, A. (2016). Using Automatic Item Generation to Improve the Quality of MCQ Distractors. *Teaching and Learning in Medicine*, *28*(2), 166–173. https://doi.org/10.1080/10401334.2016.1146608

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Matthewson, K., Tessler, M. H., Creswell, A., McClelland, J. L., Wang, J. X., & Hill, F. (2022). *Can language models learn from explanations in context?* (arXiv:2204.02329). arXiv. http://arxiv.org/abs/2204.02329

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A Paradigm Shift from “Human Writing” to “Machine Generation” in Personality Test Development: An Application of State-of-the-Art Natural Language Processing. *Journal of Business and Psychology*, *38*(1), 163–190. https://doi.org/10.1007/s10869-022-09864-6

Liu, X., Lu, H., Zhou, Z., Chao, M., & Liu, T. (2022). Development of a computerized adaptive test for problematic mobile phone use. *Frontiers in Psychology, 13*, 892387.

Margetić, B., Peraica, T., Stojanović, K., & Ivanec, D. (2022). Spirituality, Personality, and Emotional Distress During COVID-19 Pandemic in Croatia. *Journal of Religion and Health*, *61*(1), 644–656. https://doi.org/10.1007/s10943-021-01473-6

May, S., Littlewood, C., & Bishop, A. (2006). Reliability of procedures used in the physical examination of non-specific low back pain: A systematic review. *Australian Journal of Physiotherapy*, *52*(2), 91–102. https://doi.org/10.1016/S0004-9514(06)70044-7

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *ETS Res. Rep*. 16, i–30. doi: 10.1177/014662169201600206

Park, I.-J., Shim, S.-H., Hai, S., Kwon, S., & Kim, T. G. (2022). Cool down emotion, don’t be fickle! The role of paradoxical leadership in the relationship between emotional stability and creativity. *The International Journal*

of Human Resource Management, *33*(14), 2856–2886. https://doi.org/10.1080/09585192.2021.1891115

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459–467. https://doi.org/10.1002/nur.20199

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *J. Educ. Stat*. 4, 207–230. doi: 10.3102/10769986004003207

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information system (PROMIS). *Med. Care* 45, S22–S31. doi: 10.1097/01.mlr.0000250483.85507.04

Samejima, F. (1969). Estimation of latent ability using a pattern of graded responses. *Psychometrika* 34:100. doi: 10.1007/BF03372160

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat*. 6, 461–464. doi: 10.1214/aos/1176344136

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Von Davier, M. (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika*, *83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Wang, M.C., Dai, X.Y., Yao, S.Q. (2011). The Application of CAT on Emotional Intelligence with Item Response Theory. Chinese Journal of Clinical Psychology,*19(4)*,454-457. [王孟成,戴晓阳,姚树桥.(2011).中国大五人格问卷的初步编制III:简式版的制定及信效度检验.*中国临床心理学杂志*(04),454-457.doi:10.16128/j.cnki.1005-3611.2011.04.004.]

Wettstein, A., Ramseier, E., & Scherzinger, M. (2021). Class- and subject teachers' self-efficacy and emotional stability and students' perceptions of the teacher–student relationship, classroom management, and classroom disruptions. *BMC Psychology*, *9*(1), 103. https://doi.org/10.1186/s40359-021-00606-6

Xu, L., Jin, R., Huang, F., Zhou, Y., Li, Z., & Zhang, M. (2020). Development of Computerized Adaptive Testing for Emotion Regulation. *Frontiers in Psychology*, *11*, 561358. https://doi.org/10.3389/fpsyg.2020.561358

Zhang, B., Li, Y. M., Li, J., Luo, J., Ye, Y., Yin, L., Chen, Z., Soto, C. J., & John, O. P. (2022). The Big Five Inventory–2 in China: A Comprehensive Psychometric Evaluation in Four Diverse Samples. *Assessment*, *29*(6), 1262–1284. https://doi.org/10.1177/10731911211008245

Zhang, L. F., Liu, K., Song, G., & Tu, D. B. (2020). The application of cat on emotional intelligence with item response theory. *Journal of Jiangxi Normal University (Natural Science)*, *44(5)*, 454–461.[张龙飞,刘凯,宋鸽

涂冬波.(2020).计算机化自适应测验技术在情绪智力智能测评中的初步应用——基于项目反应理论.*江西师范大学学报(自然科学版)*(05),454-461.doi:10.16357/j.cnki.issn1000-5862.2020.05.02.]

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*, *160*.